Clustering Music by Genres Using Supervised and Unsupervised Algorithm

Kyuwon Kim, Wonjin Yun, Rick Kim
CS229 Machine Learning Project, Stanford University

Objectives

Most music recommender systems use either a collaborative filtering mechanism or a content-based filtering mechanism. However, both mechanisms require large amount of data. A collaborative filtering relies on data of peer users, and a content-based filtering needs prior labels of the piece of music. Our objective is to use a supervised or an unsupervised algorithm to efficiently separate a set of unlabeled music samples into groups without the help of any external data.

Data and feature sampling

Five different genres of music were chosen as our data classes: classic, EDM, hip-hop, jazz and rock, with 60 samples of music from each genre. Samples were randomly streamed from YouTube.

DFT was performed on each sample and grouped the result into certain blocks of frequencies. Average values of magnitudes in each block were used as our features. Two different feature sets X_L and X_M were used for our test. X_L used a finite subdivision in the low frequency region (10 ∼ 200Hz), while X_M uses a Mel-scale frequency division for 20 ∼ 2000 Hz. Each feature vector was normalized to satisfy $||x^{(i)}|| = 1$.

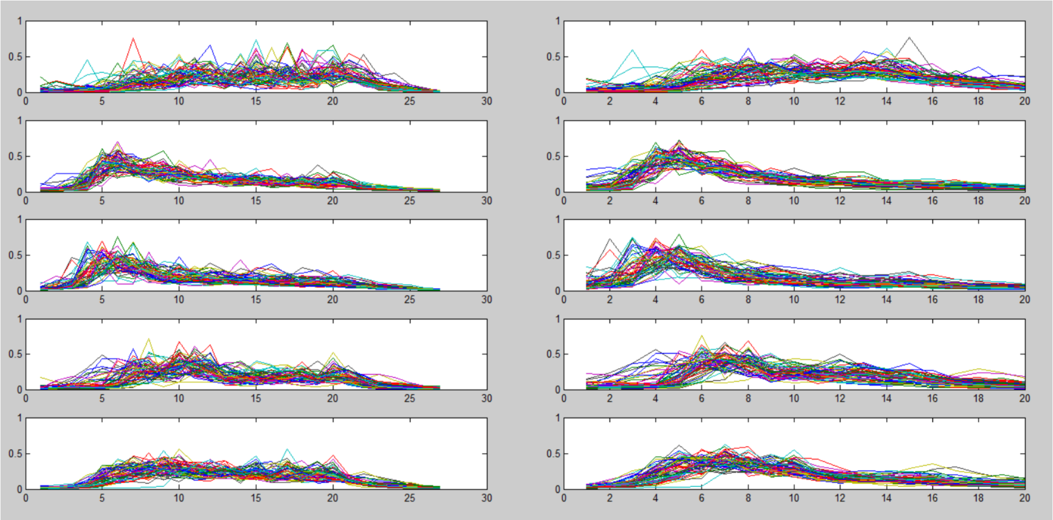

Fig.1: Features X_L (left) and X_M (right) sampled from 5 genres : classic, EDM, hip-hop, jazz and rock.

Supervised Learning

An equal amount of training data from each class was randomly selected. ***Classification and Regression Tree (CART)*** was trained on the training data and the performance of a CART classifier was evaluated on the test set ($m_{test} = \frac{1}{6} m_{data}$). For a 3 genre classification, X_M showed an accuracy of **86.7%** ($\sigma = 4.27$) compared to 77.2% ($\sigma = 7.42$) of X_L. The accuracy showed a significant drop when tested on a 5 genre classification. (**60.7%** ($\sigma = 4.32$) and 54.7% ($\sigma = 6.02$), respectively)

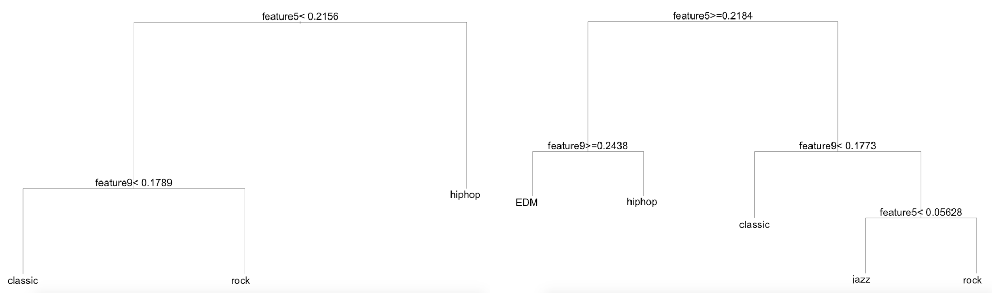

Fig.2: CART for 3 (Left:$m_{train} = 150$)) genres and 5 (Right:$m_{train} = 250$) genres are demonstrated using X_L.

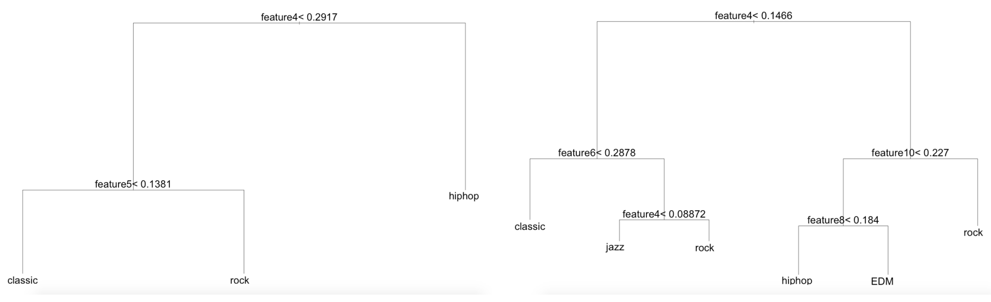

Fig.3: CART for 3 (Left) and 5 (Right) genres are demonstrated using X_M.

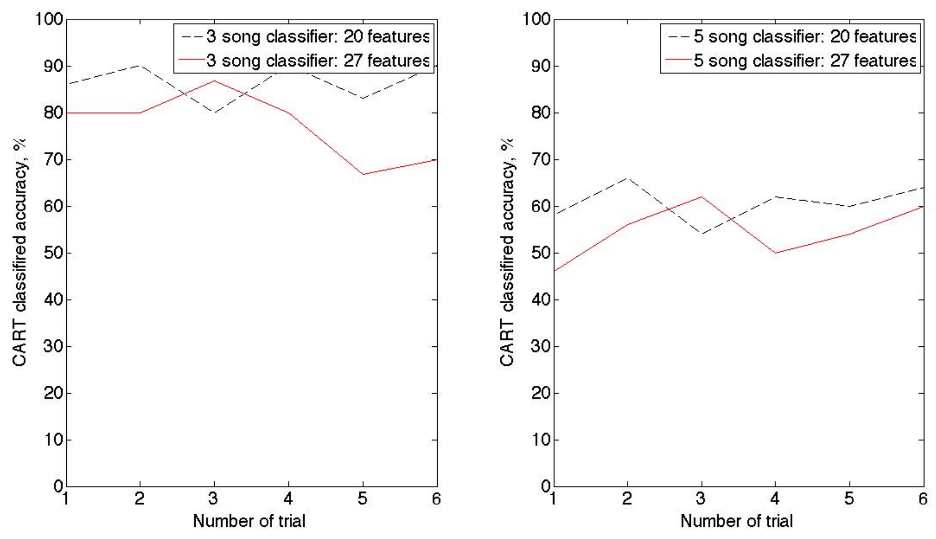

Fig.4: CART classifier accuracy test for the 6 given random selections of training data. Using 20 and 27 features, each 3 (Left) and 5 (Right) genres are classified.

Unsupervised Learning

K-means clustering was performed on our data set to cluster the samples. One sample from each genre were randomly chosen as our initial pivots, which were labeled to correctly match each cluster with its genre. Test was performed on both feature sets X_L and X_M, using 10 most significant vectors from PCA.

PCA improved the performance for feature set X_M (3 genres : 1.3%, 5 genres : 2%) while feature set X_L did not show a significant improvement. For a 3 genre classification, X_M showed a slightly better performance with **84.4%** accuracy. Meanwhile, in the 5 genre classification, X_L showed a better performance with an accuracy of **62.0 %**.

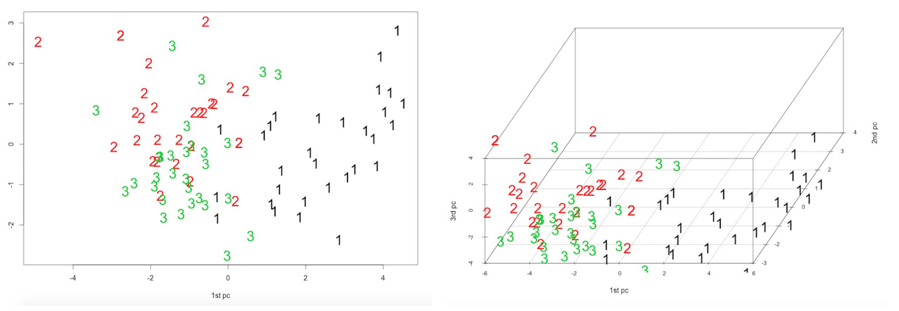

Fig.5: PCA plot from 3 genres effectively visualizes the distance between songs. The plot (left) with first two principal components efficiently classifies classic (1) from the other genres , but three principal components is even unable to effectively separated the hip-hop(2) and rock (3) indicating the requirement of better feature selection.

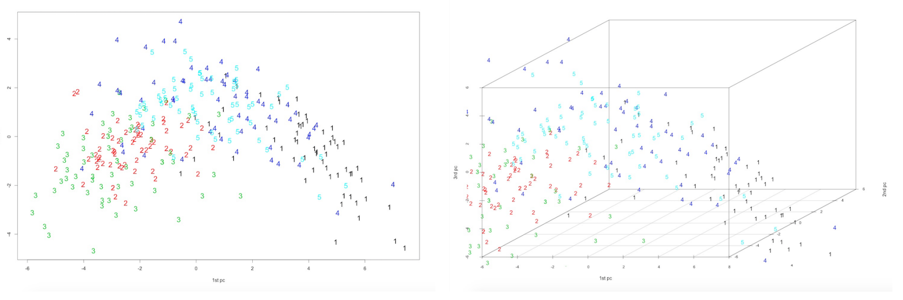

Fig.6: PCA plot from 5 genres effectively visualizes the similarity and difference between songs. The plot (left) with first two principal components exhibits the genre similarity between EDM (2) and hip-hop (3); jazz (4) and rock (5). The plot (right) with three principal components demonstrates the widest span of jazz (4) compared to the highly-localized classic (1).

	C	H	R	Acc.
C	45	0	15	75.0
H	0	49	11	81.7
R	5	1	54	90.0
Acc.	90.0	98.0	67.5	82.2

	C	H	R	Acc.
C	50	0	10	83.3
H	0	51	9	85.0
R	5	4	51	85.0
Acc.	90.9	92.7	72.9	84.4

Fig.7: K-means result for 3 genres using X_L (left) and X_M (right)

	C	E	H	J	R	Acc.
C	37	0	0	9	14	61.7
E	0	51	7	0	2	85.0
H	0	30	26	0	4	43.3
J	3	7	1	39	10	65.0
R	3	16	0	8	33	55.0
Acc.	86.0	49.0	76.5	69.6	52.4	62.0

	C	E	H	J	R	Acc.
C	47	0	2	9	2	78.3
E	0	48	5	0	7	80.0
H	0	35	17	0	8	28.3
J	10	2	2	23	23	38.3
R	4	6	0	6	44	73.3
Acc.	77.0	52.7	65.4	60.5	52.4	59.7

Fig.8: K-means result for 5 genres using X_L (upper) and X_M (lower)

Distance between genres

The learning algorithms showed significantly better performances for a 3 genre classification. The result from Fig.8 shows that EDM was frequently confused with hip-hop, while jazz was often mistaken as rock. The PCA result from Fig.6 also demonstrates how the genres are distributed on the feature space. In order to examine the distance between genres, we constructed a neighbor graph based on the L2-norm distances of feature set X_M (Fig.9 (left)). Two songs were labeled as neighbors only if the distance is below a certain threshold. Fig.9 (right) shows a distance score heatmap constructed using the graph, where the score was defined as :

$$s_{ij} = \sum_{k \neq i,j} \exp(-||x^{(i)} - x^{(k)}|| - ||x^{(j)} - x^{(k)}||)$$

High score implies that the two data has common neighbors and are likely to be classified as same genre. The heatmap shows some genres that are likely to be mistaken.

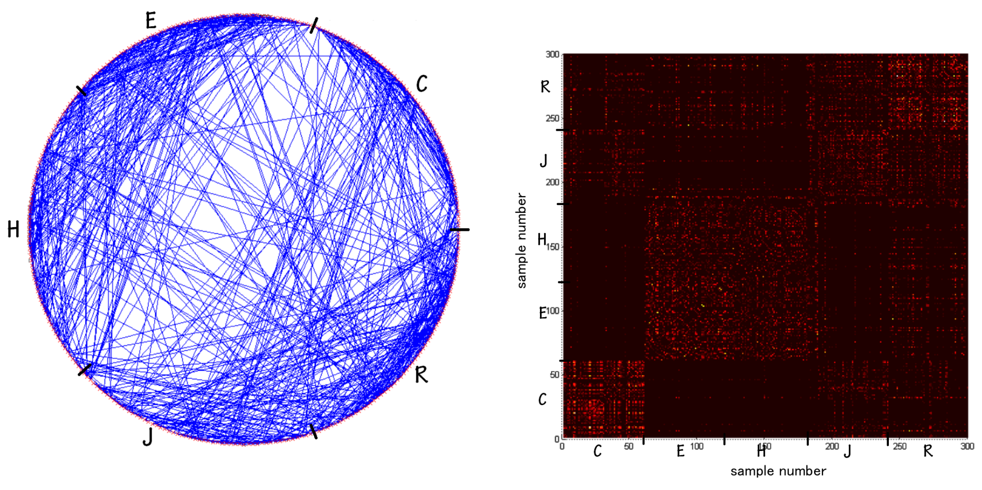

Fig.9: (left) Graph connecting songs with small feature distance. Each red dot represents a song. (right) Heatmap of a distance matrix showing classic(1-60), EDM(61-120), hip-hop(121-180), jazz(181-240) and rock(241-300).

Conclusion

- Using a supervised learning algorithm (CART), efficient data examination and corresponding feature extraction were successfully performed.
- Supervised and unsupervised (Cross-Validation & PCA) learning provided the efficient examination and visualization of the property of music data showing the genre similarity and discrepancy based on the principal components.
- Feature extraction corresponding to the data examination were successfully performed resulting in two sets of frequency blocks using X_L and X_M.
- For Supervised learning, CART showed a classification accuracy of **86.7%** for 3 genres and **60.7%** for 5 genres at maximum.
- K-means showed a classification accuracy of **84.4%** for 3 genres and **62.0%** for 5 genres at maximum.

References

Soltau, H., Schultz, T., Westphal, M., Waibel, A. (1998). *Recognition of music types.* In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference* on (Vol. 2, pp. 1137-1140). IEEE.

Shao, X., Xu, C., Kankanhalli, M. S. (2004). *Unsupervised classification of music genre using hidden markov model.* In *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference* on (Vol. 3, pp. 2023-2026). IEEE.

Tsai, W.H., Bao, D.F. (2010). *Clustering music recordings based on genres.* In *Information Science and Applications (ICISA), 2010 International Conference* on (pp. 1-5). IEEE.